RESEARCH ARTICLE OPEN ACCESS

Discovering diamonds under coal piles: Revealing exclusive business intelligence about online consumers through the use of Web Data Mining techniques embedded in an analytical customer relationship management framework

Myriam Ertz, M.Sc*, Raoul Graf, Ph.D**
*(Department of Marketing, School of Management Sciences, University of Quebec in Montreal, Canada)
** (Department of Marketing, School of Management Sciences, University of Quebec in Montreal, Canada)

ABSTRACT
Web Mining has gained prominence over the last decade. This rise is concomitant with the upsurge of pure players, the multiple challenges of data deluge, the trend toward automation and integration within organization, as well as a desire for hyper segmentation. Confronted, partly or totally, with these multiple issues, companies recourse increasingly to replicate the data mining toolbox on web data. Although much is known about the technical aspect of WM, little is known about the extent to which WM actually fits within a customer relationship management system, designed at attracting and retaining the maximum amount of customers. An exploratory study involving twelve senior professionals and scholars indicated that WM is well-suited to achieve most of the customer relationship management objective, with regards to the profiling of existing web customers. The results of this study suggest that the engineering of WM processes into analytic customer relationship management systems, may yield highly beneficial returns, provided that some guidelines are scrupulously followed.
Keywords: Web mining, analytical customer relationship management, segmentation, data mining, profiling

I. INTRODUCTION

With the advent of big and large data originating from the Internet (e.g. social media, geo-referencing, credit card information), the World Wide Web is a large and ever-growing database, which constitutes a fertile area for analytical research, but does also require increasing work on data carpentry and online analytical systems engineering [1].

In addition to overwhelming data flows, automated, flawless and integrated processes tend to become the norm in the industry since businesses tend to rely heavily on integrated processes (e.g. Customer Relationship Management Systems, or Enterprise Resource Planning Systems), to manage every aspects of customer relationships [2]. Those systems are also increasingly more fed with web data (Web Houses) [2].

Increased and diversified data, as well as more integrated analytical systems, enable companies to craft hyper-segmentation strategies. The level of targeting becomes as low as that of the individual, with markets being fragmented into micro-segments [3]. The level of targeting narrowed from mass markets to unique individuals. Meanwhile, traditional research (e.g. surveys) is time-consuming, costly and bias prone, it is also increasingly more difficult to administer to increasingly busy consumers [4, 5].

Given the multiple challenges of data deluge, automation and integration trends, as well as hyper-segmentation, that companies increasingly face, is web mining a solution? More specifically, is it a useful approach to sift through the large data about existing online customers, and integrate that knowledge into Customer Relationship Management (hereafter, CRM) systems to carve-craft powerful personalized strategies?

More specifically, this study seeks to answer the following research questions: (1) to what extent do WM methods applied to web data provide accurate profiles of existing web customers? (2) To what extent do WM methods applied to web data identify strategically important existing web customers? (3) To what extent do WM methods identify existing web customers' loyalty or defection statuses?

In this study, we investigate the extent to which Web-Mining (hereafter, WM) within an analytical CRM framework, profiles well existing customers who interact with a given E-commerce platform.

II. LITERATURE REVIEW

1.1. Web mining

WM refers to the automatic discovery and extraction of information from web data [6,7]. WM refines large, complete, integer, reliable and cheap data into a time-effective manner [4]. It therefore draws on the vast web data and overcomes traditional market research drawbacks. WM enables one-to-one relationships and thus mass customization [8], so that it may contribute to achieve hyper-segmentation. Besides, if appropriately disseminated throughout organizational layers and divisions, WM is a valuable part of analytical CRM (a CRM), acting as a powerful platform for tactic and strategic decision-making [9].

However, despite its multiple advantages, WM is hard to implement from an operational viewpoint [10]. Second, for an improved dissemination of WM extracted knowledge across the organization, WM needs to be integrated into the CRM applet of the marketing function as well as into the broader Knowledge Management (KM) or Business Intelligence (BI) framework of the entire organization, which requires typically tremendous business process re-engineering. Third, web data are generally very large and not always meaningful leading to poor data quality issues if not appropriately sifted through [11]. Fourth, WM should not be a standalone technique randomly appended to the already existing arsenal of data analytics techniques. Rather, companies that adopt market-oriented strategies to better compete on the global marketplace, need to assess objectively and critically the benefits of the WM methods and techniques and their place as well as utility as inputs of the organizational decision-making process. Eventually, an important challenge for organizations, is to be able to diffuse and disseminate efficiently the information derived from WM by being holistically integrated to the aCRM applet of CRM systems and to KM-BI systems.

1.2. Analytical Customer Relationship Management Systems

Xu and Walton [12] developed a typology of he four main objectives of a CRM-enabled customer knowledge acquisition framework: (1) profiling existing customers; (2) Explaining the behavior of existing customers; (3) profiling prospective customers, and (4) Explaining the behavior of prospective customers. The framework refers however to an offline-based CRM process encompassing Data-Mining, forecasting, and scoring techniques [2]. Data from the web, and thus WM techniques to process them, are not comprised in this framework. This study focuses on evaluating the use of WM methods and techniques to fulfill the first objective of Xu and Walton's [12] typology of profiling existing customers.

By drawing on Xu and Walton's [12] framework, it is possible to develop a framework of a CRM framework for web users' knowledge acquisition. The overall assumption is that, integrating the WM process into a CRM framework is assumed to turn operational web data into meaningful and relevant knowledge of current web customers.

1.3. Web Mining techniques for profiling existing web customers

In order to profile existing web customers, WM offers a many useful methods and their respective techniques:

- Clustering method: creating homogeneous groups of customers (e.g. agglomerative or hierarchical clustering, K-means, TwoStep clustering, Kohonen network/Self-Organizing Map, K-nearest neighbour, principal component analysis, factor analysis) [13];
- Classification method: explain or predict the qualitative characteristic of an individual based on other qualitative or quantitative characteristics of that individual (e.g. decision trees, artificial neural networks, discriminant analysis, logistic regression, decision rules, support vector machines, Bayesian networks) [14,15];
- Prediction method: explain or predict the quantitative characteristic of an individual based on other quantitative characteristics of that individual (e.g. decision trees, artificial neural networks, ordinary least squares regression, support vector machine, generalized linear models) [14].

Clustering techniques provide clusters of customers, which may be used to reorganize the website according to discovered clusters, as well as models to enrich the database by integrating the code of the cluster in the customer database or by integrating the model directly on the website [13]. *Classification techniques* provide scores that enrich the database since the scores are integrated in the customer database; and models that are useful to develop recommendation modules (e.g. intelligent agents, recommendation systems, choice matrices) [1]. Therefore, the first research propositions read as follows:

RP1.1: web data generated by existing web customers are sufficiently detailed and accurate to provide a strong basis for the creation of precise profiles about existing web customers.

RP1.2: Clustering and classification techniques applied to web data (e.g. web log data, search results, and web pages) create homogeneous groups of existing web customers.

Classification and prediction methods provide models that enrich the database by integrating the score in the customer database for future predictions [14]. One important variable on which customers are generally classified refers to the Recency, Frequency, and Monetary (RFM) of a customer's purchases, which assesses her individual value in the form of the Customer Lifetime Value (CLV) score [16]. RFM-based CLV and profit-cost ratios identify strategically important customers in a web context. According to past research, classification and prediction techniques are well-suited to assess CLVs based on RFM data [17]. Therefore, the next research propositions posit that:

RP2.1: Web data generated by existing web customers encompass enough information about the profit-cost, the RFM of purchases made by existing web customers, which contributes to identify strategically important existing customers.

RP2.2: Classification and prediction methods applied to web data predict the value of a given web customer to identify strategically important existing web customers.

Web Usage Mining (WUM) refers to a global analysis that records the user's behavior and how she interacts with an application from the instant she accesses a site to the moment when she leaves the site [14]. Besides, Web Content Mining (WCM) refers to the analysis of the content of web pages, whereas Web Structure Mining (WSM) refers to the analysis of links and hyperlinks. Through these different approaches, core classification and clustering methods can be useful tools to determine important customer metrics such as loyalty and attrition statuses of one or several customers in an attempt to categorize them. The third set of research propositions reads therefore as follows:

RP3.1: Web data generated by existing web customers indicate whether a web customer is loyal to a given business or defects from that business.

RP3.2: Classification and clustering methods applied to web data predict membership of an individual to the loyal or defecting customer group.

III. METHODOLOGY

A questionnaire was developed based on the research propositions to be explored. Since the design of the study is highly exploratory, the questions were exclusively written in an open-ended form. A convenience sample was drawn from a pool of potential participants. A total of twelve valid in-depth semi-structured interviews were conducted. A condition of eligibility was that participants had to have a thorough knowledge of WM methods and techniques and also a sound understanding of business issues. About half of the respondents were senior directors or c-level executives in public or private organizations, while the other half consisted mainly of scholars from IT, IS, marketing, statistics, mathematics or engineering disciplines. The sample is therefore heterogeneous enough to allow for a diversity of opinion and responses. Respondents input was tape-recorded, transcribed and analyzed using a response matrix, in which each participant is crossed with each research theme, corresponding to the six research propositions.

IV. RESULTS

Both research propositions 1.1 and 1.2 appear valid. WM methods applied to web data (e.g. web log data) provide accurate profiles of existing web customers. More specifically, in response to RP 1.1, internal company data may be coupled with external syndicated data. Both should be large, granular, of good quality, ideally issued by logged in web users. Regarding RP1.2, both online and offline data should be triangulated to optimize the segmentation of existing web customers. Besides, recent advances in WM have made WM tools directly actionable on the website. Visitors are segmented "on the spot" and subsequent personalization allows for immediate and dynamic customization. Fig. 1 summarizes he findings related to RP1.1 and RP1.2 (answering the first research question).

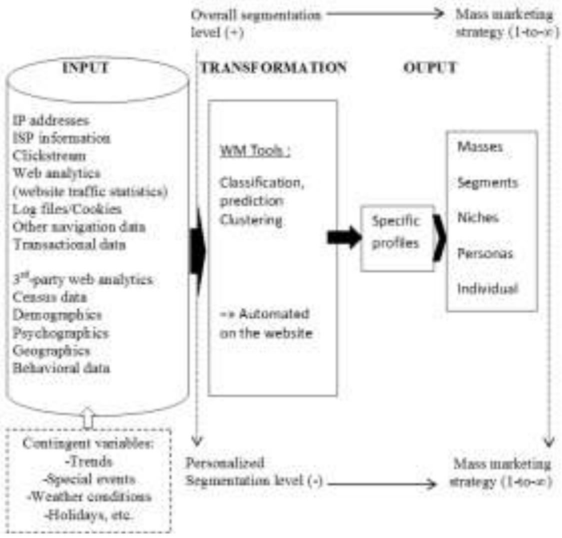

Figure 1. WM-enabled process of profiling existing web customers

Research propositions 2.1 and 2.2 are validated. WM techniques determine successfully web customers' RFM, CLV, and thus, strategic importance. Additional specificities need however to be met regarding web data adequacy (cf. RP2.1) since the RFM-based computation of CLV requires transactional data that is easily accessible. Besides, CLV computed with offline data can enrich the companies' insight into consumers' specific behavior in the future and supports the crafting of relevant E-business strategies to maximize returns. Regarding RP2.2, business rules can be derived by aggregating huge datasets of CLV-based segmentations that were obtained by means of WM techniques. Fig. 2 summarizes the key findings pertaining to the second blog of research propositions (in answer to the second research question).

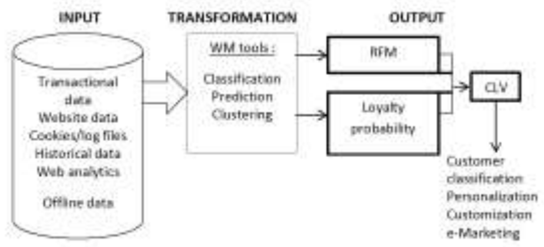

Figure 2. WM-enabled identification of strategically important existing web customers

Eventually, the findings lend support to RP3.1 and RP3.2 so that WM techniques enable companies to identify existing web customers' loyalty as well as defection statuses, which may constitute an additional segmentation approach. With respect to RP3.1, data preprocessing (i.e. filtering, selecting, cleansing, formatting) is of utmost importance, especially for this specific objective of loyalty and defection status identification. Regarding RP3.2, automated methods should be preferably used in order to compute, in real-time, the loyalty or defection status (risk or extent) of the customer. Such knowledge constitutes then an opportunity to customize web page content or structure to retain consumers who are identified as being likely to defect; or an opportunity to display cross-/up-/deep-selling content for consumers who are likely to remain loyal. Multiple other forms of surgical content engineering strategies may therefore be implemented. Fig. 3 summarizes the findings pertaining to the third research question.

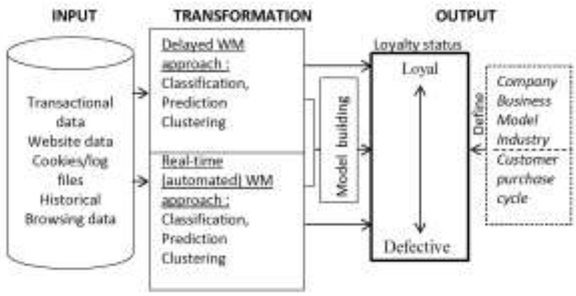

Figure 3. WM-enabled identification of existing web customers' loyalty or defection statuses

As a wrap-up, Fig. 4 summarizes the findings related to the six research propositions altogether. At the core of the figure, WM enables to classify existing web customers into company-specific categories, and therefore the consumer varies from a business-specific low to high profile. The variables which can be used to determine the extent to which a consumer may exhibit more or less of a business-specific profile, may be done in a number of ways. In this study, we explored two approaches, which are frequently used in business contexts: (1) the RFM-based CLV approach, and (2) the loyalty vs. defection probability. Others may exist though. Both loyalty status (ranging from defection to loyalty likelihood) and customer strategic importance (ranging from unprofitable to profitable), may therefore be complemented by many other attributes and variables to form multi-attribute profiles. As with most data analytic approaches, the variable used to determine an existing web customers' business-specific profile determines the WM-derived model, typology or rule which then also predicts the customers' future positioning on this variable. Finally, it is also posited that the more general the level of analysis, and thus the recourse to mass marketing, the more likely web consumers will defect from the company and be unprofitable. The usage of WM enables to reach a personalization level through granular marketing which is strongly associated with more profitable and loyal web consumers.

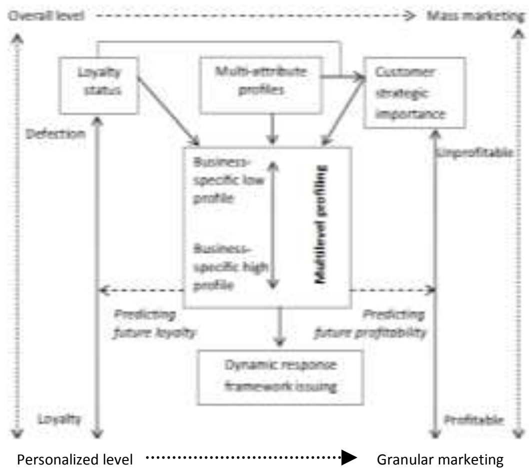

Figure 4. WM-enabled process of existing web customers profiling

V. DISCUSSION

Depending on the level of segmentation sought by the company and, therefore, the strategic marketing orientation, resulting from the subsequent tactical (i.e. targeting) efforts deployed to implement those strategies, WM has several benefits. First, it identifies the profile of existing web customer on a multitude of attributes. Two of such attributes were investigated in the current study. The findings revealed that, WM determines well the loyalty status (defection vs. loyalty) of existing web customers; and finally, WM evaluates equally well the strategic importance of existing web customers (unprofitable vs. profitable).

WM enables to fulfill the three first objectives of a CRM, as delineated by Xu and Walton [12]. Knowledge about these three aspects is useful in order to develop business-specific profiles ranging on a continuum from low to high, or even in categorical nature, given the segmentation type being sought by the company.

VI. MANAGERIAL IMPLICATIONS

WM-powered CRM processes enable successfully to predict existing web customers' future behavior and to develop dynamic response framework in real-time that are appropriate to the current as well as future customer's profile.

Yet, managers should remain conscious of the caveats that surround WM. First, practitioners should be aware of the "garbage in/garbage out" issue since all information is not necessarily good or useful to process. An important job of comparison, judgement and selection needs to be done prior to conduct any WM project. Second, according to most respondents, data produced by logged in customers appear the best kind of data since profiles can be confidently attached to a specific customer. Websites may therefore benefit from implementing a technology of user recognition which logs automatically the customer in, without her having to do anything. Third, companies' databased should also allow for high volumes of data entries, since higher quantities of data leverage better results. Fourth, in addition to data quality, special care should be addressed to the planning of the WM project, as well as to the robustness of the WM analytical process. Adherence to such principles may prove to yield the higher returns on investments.

VII. CONCLUSION

This study positions itself into the growing literature stream exploring WM engineering for business purposes [4, 13, 14, 15, 17, 18, 19]. In that respect, the study investigated the benefits of using WM methods and techniques in order to reach the specific aCRM objective of profiling customers in an online context.

The authors considered WM methods from a broad viewpoint, without going into the particularities of each WM technique. The WM field of research is also fast-evolving. Therefore, new methods and techniques may have emerged. Additional research could focus on investigating broader arrays of WM methods and techniques and also investigate some WM techniques in particular.

We only considered Xu and Walton's [5] aCRM framework, but other aCRM models or typologies may exist. It would therefore be of interest, for future research, to investigate the extent to which WM fits into other CRM frameworks, in general, and aCRM ones, in particular.

Finally, a particularly promising avenue of research is being paved by the recent technological advances in sentiment analysis and opinion mining, both being conflated with WM. Beyond the transactional data-which were studied in this article-the web should also be mined for feelings [17]. This study remained essentially focused on facts. Yet, future research could determine the extent to which sentiment analysis may be usefully enacted in order to create additional segmentation variables within the already existing business specific classificatory scheme. For example, conversations, opinions, sentiments, and other content-related variables may enable companies to enrich their existing fact-based classification schemes with variables that are more attitudinal and emotional in nature. WM could contribute tremendously to further knowledge in that regards.

VIII. Acknowledgements

The authors are thankful to the Canadian Marketing Association (CMA) as well as to the Marketing Research and Intelligence Association (MRIA) for their kind support in our recruitment of the participants for this study.

REFERENCES

[1] L. Van Wel, and L. Royakkers, Ethical issues in web data-mining, *Ethics and Information Technology, 6,* 2004, 129-140.

[2] R. Kimball, and M. Ross, The data warehouse toolkit: The complete guide to dimensional modeling (2nd Ed.) (New York, NY: Wiley and sons, 2002).

[3] J. Bousquet, Y. Lachance, S. Laferté, and F. Marticotte, *Marketing stratégique* (Québec : Chenelièere Éducation, 2007).

[4] I. Mihai, *Web-Mining in E-commerce*, Annals of the University of Oradea, Economic Science Series. 959-962, 2009.N. K. Malhotra, *Marketing research: An applied orientation* (6th ed.) (Upper Saddle River, NJ: Pearson Education, 2010).

[5] R. Cooley, B. Mobasher, and J.Srivastava, Web mining: Information and pattern discovery on the world wide web, *Proc. 9th IEEE International Conf. on Tools with Artificial Intelligence*, Newport Beach, CA, 1997.

[6] R. Kosala, H. Blockeel, Web mining research: A survey, *Proc. ACM SIGKDD Explorations*, *2(1)*, 2000, 1-15.

[7] B. Liu, B. Mobasher, and O. Nasraoui, Web usage mining in B. Liu (Ed.), *Web data mining: Exploring hyperlinks, contents, and usage data*, (Berlin: Springer-Verlag, 2011) 527-603.

[8] J. Ranjan, V. Bhatnagar, Role of knowledge management and analytical CRM in business data mining, *The Learning Organization*, *18(2)*, 2011, 131-148.

[9] Aggarwal, V. Mangat, Application areas of web usage mining, *Proc. 5th International Conference on Advanced Computing & Communication Technologies (ACCT)*, Haryana, India, 2015, 208-211.

[10] A. Reid, M. Catterall, Hidden data quality problems in CRM implementation in H. E. Spotts (Ed.), *Marketing, technology and customer commitment new economy*, (Springfield, IN: Springer International Publishing, 2015) 184-189.

[11] M. Xu, and J. Walton, Gaining customer knowledge through analytical CRM, *Industrial Management and Data Systems, 105(7)*, 2005, 955-971.

[12] M. Bazsalicza, and P. Naim, *Data mining pour le web: Profiling, filtrage collaborative, personalisation client* (Paris: Eyrolles, 2001).

[13] F. Tufféry, *Data-mining et statistiques décisionnelles: L'intelligence des données* (Paris : Technip, 2011).

[14] K. J. Cios, R. W. Swiniarski, L. A. Kurgan, Data Mining, (New York, NY: Springer, 2007).

[15] M. Jeffery, Data-driven marketing: The 15 metrics everyone in marketing should know (Evanston, IL: Northwestern University Press, 2010).

[16] A. lbadvi, and M. Shahbazi, Integrating rating-based collaborative filtering with customer lifetime value: New product recommendation technique, *Intelligent Data Analysis*, *14,* 2010, 143-155.

[17] A. Wright, Mining the web for feelings, not facts, *The New-York Times,* August 23, 2009.

[18] M. Ertz, and R. Graf, How do theybehave on the web? An exploratory study of mining web for analytical customer relationship management, *International Journal of Electronic Commerce Studies, 6(2),* 2015, 289-304.

[19] M. Ertz, R. Graf, Spotting the 'elusive' prospect customer: Exploratory study of a web-powered customer relationship management framework, *Journal of Applied Business Research, 31(5),* 2015, 1935-1850.